PREDICTING CRASH OCCURRENCE AT INTERSECTIONS IN TEXAS: AN OPPORTUNITY FOR MACHINE LEARNING

Theodore Charm
Graduate Research Assistant
Department of Government
The University of Texas at Austin
theodorecharm@gmail.com

Haoqi Wang
Graduate Research Assistant
Department of Biomedical Engineering
The University of Texas at Austin
haoqiwang@utexas.edu

Natalia Zuniga-Garcia
Research Fellow
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin
nzuniga@utexas.edu

Mostaq Ahmed
Department of Community and Regional Planning
The University of Texas at Austin
mostaq@utexas.edu

Kara M. Kockelman
(Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin
kkockelman@mail.utexas.edu

Word Count: 6,446 words + 2 Tables (250 words per table) = 6,946 word-equivalents

July 30, 2022

¹These authors contributed equally.
ABSTRACT

This paper studies the frequency of traffic crashes at intersections across Texas by employing zero-inflated negative binomial (ZINB) models using the MLE method, and various tree-based ML methods, namely random forests (RF), XGBoost, LightGBM, and Bayesian additive regression trees (XBART) to predict the frequency of crashes at intersections. Official records of traffic crashes from 2010 to 2019 were used in addition to the roadway inventory database and other sources to map more than 700k intersections. The performances of the MLE and ML models were computed and compared, using R-square and Root Mean Square Error as the metrics. Results indicated that RF had the best model performance in predicting crash frequency. Resampling the data led to better prediction performances for all the models and was useful in dealing with highly imbalanced crash data. Road design variables had the highest feature importance on the ML models to predict crash occurrence. Sensitivity analysis showed that the effects of several predictors have different directions across different ML models making interpreting their contribution in predicting crash occurrences difficult. The findings suggest that machine learning models are better at predicting crash occurrences, whereas statistical models are better at investigating the contributing factors of a crash event.

Keywords: Motor vehicle crashes, intersection safety, machine learning, deep learning, crash counts

BACKGROUND

Traffic crashes are very expensive and often society has to pay in human lives. Motor vehicle crashes are one of the leading killers in the U.S., with over 100 deaths every day (National Center for Statistics and Analysis, 2017). In 2015, more than 2.5 million Americans were taken to emergency departments due to injuries sustained in a motor vehicle crash (CDC, 2018). Economically speaking, in 2017 in the U.S., the cost of medical care, loss of productivity, loss of lives, etc. - all sum up to more than $75 billion (CDC, 2018).

Although there is a significant amount of research in predicting motor vehicle crashes, there is still a lot to explore to gain a better understanding of pre-crash conditions and return a more accurate prediction, which is crucial for pro-active road-safety management.

The majority of the existing literature on crash frequency prediction modeling adopted econometric modeling approaches (Lord and Manning, 2010; Yasmin and Eluru, 2018; Wang et al., 2018). Dionne et al. (1993) and Jovanis and Chang (1986) argued in favor of the Poisson regression model in their study relating exposure variables to crash counts. In an attempt to develop crash prediction modeling for Italy, Caliendo et al. (2007) investigated the comparative suitability of the Poisson, Negative Binomial, and Negative Multinomial distributions. They found the over-dispersion of the crash data was making the Poisson model particularly weaker than the negative binomial model. Another problem with crash data is the presence of a lot of zeros in the dataset indicating the locations that have zero crash events. Several studies found that due to unobserved heterogeneity and the presence of excess zeros, the zero-inflated negative binomial model performed better than the regular negative binomial model. These econometric models often fail when working with complex and highly nonlinear motor vehicle crash data (Karlaftis and Vlahogianni, 2011). To deal with these limitations of statistical models, several machine learning techniques, including decision tree-based models, Artificial Neural Network (ANN), Support Vector Machine (SVM), and deep learning models, have been applied to various traffic crash prediction models because they do not heavily rely upon the certain types of underlying assumptions and relationships between the dependent variable and the contributing factors (Dong et al., 2018; Rahman, 2018). Among many machine learning techniques, tree-based models are being widely used in traffic safety literature because of their capability to identify the complex pattern of crash likelihood and their interpretability in explaining the relationship between target variables and the predictor variables (Chang and Chen, 2005; Chang and Chien, 2013; Rahman, 2018; Zuniga-Garcia et. Al., 2021). In another study, Chang (2005) compared the prediction accuracy of the negative binomial regression model with ANN in crash frequency prediction and found that ANN offers higher accuracy as a consistent alternative to the NB model. Dong et al., (2018) developed a deep learning model with a multivariate negative binomial (MVNB) regression layer and reported that the model provides better traffic crash prediction across different levels of injury severity.
To address the heterogeneity of the crash data, some studies applied data clustering methods prior to applying machine learning models for crash prediction and reported the effectiveness of clustering treatment for most cases (De Oña et al., 2013; Eluru et al., 2012; Kaplan and Prato, 2013; Zhao et al., 2019). Many previous studies used accuracy or a loss function optimized for accuracy as the validation tool to measure the performance of the crash prediction model (Abdelwahab and Abdel-Aty, 2001; Chang and Wang, 2006; Chen et al., 2016; Z. Li et al., 2012; Yu and Abdel-Aty, 2013; Zheng et al., 2019). The problem with using only the prediction accuracy to evaluate the model is that it can be misleading because of the highly imbalanced traffic crash data (Rahim and Hassan, 2021; Guo et al., 2008). Accuracy puts higher weight on the common class in an imbalanced dataset which leads to poor performance for rare classes like fatal crashes. A number of recent studies include precision and recall metrics to handle the imbalanced data problem which penalizes the model for discounting the rare classes (Jeong et al., 2018; Elamrani Abou, 2020; Elassad et al., 2020; Fiorentini and Losa, 2020). A high precision and recall value for a class means the model treated the class properly, and a low value means the model treated the class poorly.

Prediction accuracy was not the main focus of the study utilizing statistical models in traffic safety literature; the main focus was to investigate the contributing factors of crash events and different levels of crash severity (Iranaitalab and Khattak, 2017; Rahim and Hassan, 2021). The prediction performance was used mainly for validation purposes. On the other hand, machine learning models are mostly used as prediction tools in the traffic safety literature with higher accuracy and less interpretability than statistical models.

In this era of machine learning and deep learning, many new and useful techniques are still less explored in crash prediction studies. Besides, the majority of the crash prediction literature focused on road segments and only a limited number of studies focused on intersection crashes, though intersections have complex geometry and are therefore very crucial from a traffic safety perspective. Moreover, very few previous studies considered land-use and demographic variables in predicting crash frequency and severity. Hence, this article aims to contribute to the search for a better, more accurate approach to predicting crash events at intersections and to understanding the contributing factors by comparing a list of Maximum Likelihood Estimation (MLE), and Machine Learning (ML) methods. Since over 70% of the intersections had 0 crashes recorded, this paper used a zero-inflated negative binomial model estimated by maximum likelihood estimation. To handle the highly imbalanced crash data, the dataset has been resampled by implementing the \texttt{ovun.sample} function of the \texttt{ROSE} package in R, which is a “bootstrap-based technique that helps the task of binary classification in the presence of rare classes”. To improve the model performance in predicting crash counts, three types of deep learning models have been developed and evaluated. The first model has been developed using the aggregated data from 2010 to 2019, the second model has been developed to predict crash counts by severity levels and the third model used a deep neural network integrating a gated recurrent unit (GRU) network to address the longitudinal nature of crash events.

**DATA**

Crash records from 2010 to 2019 were acquired from the Texas Department of Transportation (TxDOT) Crash Records Information System or “CRIS” (TxDOT, 2020). The CRIS system collects crash reports occurring on public roadways across all 254 Texas counties, as recorded by the police. To appreciate network-level design and use information, this paper also employs the TxDOT Roadway Inventory database. The CRIS crash records were spatially matched with local land use, population, employment, household income, age, precipitation (snow and rain), and other details (like distances to the nearest hospital or school, and transit stop density). Census tract-level variables were obtained from the American Community Survey dataset (ACS, 2020). The 2015-2019 ACS 5-year estimates were used in the analysis. This paper also used annual rainfall data (1981 to 2010) from the Texas Water Development Board (2014) to obtain county-level average yearly precipitation.
Total crash counts over the recent 10-year period for all Texas intersections were obtained. Among all intersections, 522,933 (74%) had 0 crashes recorded over the 10-year period, 19.7% had 1 to 10 crashes, 2.9% had 11 to 20 crashes, and fewer than 4% had 21 or more crashes. Figure 1 illustrates the distribution of the crash counts.

The association between crash counts and a number of explanatory variables is illustrated in Figure 2, in particular, annual average daily traffic (AADT), signalized intersection, and number of lanes at major approach. The sum of AADTs for the major and minor approaches was computed, and the crash counts against the sum of AADTs is plotted in Figure 2a, which shows that intersections with frequent crashes tend to have higher-than-average AADTs. Figure 2b shows that most intersections with no signals had very few crashes. Nonetheless, a high proportion of signalized intersections had a relatively high number of crashes. In particular, about 70% and 40% of signalized intersections had more than 20 crashes and 50 crashes from 2010 to 2019, respectively. Figure 2c illustrates that with the increase of number of lanes crash counts at the intersections increase. For example, most intersections with 1 or 2 major lanes had no crashes, but about 37% of intersections with 5 to 6 major lanes had 21 or more crashes. About 40% of intersections with 7 to 8 major lanes had more than 50 crashes. Figure 2 provides evidence that intersection crashes are positively correlated with AADT, signalized intersections, and number of lanes at major approaches.
2. (a) Scatter plot for crash counts vs AADT

2. (b) Percentage of intersections by crash count range vs signalized and unsignalized intersections

2. (c) Percentage of intersections by crash count range vs lane count

Figure 2. Crash counts by AADT, presence of traffic signal and number of lanes
This paper also provides visualization of crash counts at the census-tract level. Figure 3a locates the intersections that had more than 100 crashes over the 2010-2019 period (i.e., more than 10 crashes per year on average). Such intersections are represented by the red dots in the map. Figure 3b illustrates the number of crashes per capita for each census tract. Crashes per capita were computed by dividing the total number of crashes by the population within the census tract. The percent ranks for each value are represented by different colors. Higher percent ranks are closer to the yellow end of the color spectrum, while lower percent ranks are closer to the purple end. The yellow spots are concentrated in large cities in Texas, where the census tracts have higher population densities. That indicates that large cities are more likely to have higher average crash counts than their rural counterparts. Specifically, this suggests there is an association between population density and crash counts at the census tract level. To better capture this relationship, this paper examines the Austin metropolitan area as an example, which includes the counties of Bastrop, Burnet, Caldwell, Hays, Travis, and Williamson. Figure 3c illustrates that the more densely populated census tracts tend to have higher average crash counts, in particular Travis County and the areas along the IH-35 corridor. In the next section, statistical models were used to study the association of crash counts with the explanatory variables.
REGRESSION MODEL

As a baseline, zero-inflated negative binomial models were calibrated to appreciate the effects of various explanatory variables on the total (10-year) crash counts at each of Texas’ 707,161 intersections. Since over 70% of intersections had 0 crashes recorded, this paper used a zero-inflated count model. As the standard deviation of the outcome was significantly higher than the mean, the data was over-dispersed. In light of this issue, this paper used a negative binomial model instead of a Poisson model. Therefore, zero-inflated negative binomial models were employed for regression analysis. This paper presents the full model, which included all explanatory variables in the analysis. Since the variables were measured on different scales, this paper standardized all explanatory variables to make the values comparable. First, the means were subtracted from the original values. Then, the resulting values were divided by the standard deviation, to obtain the standardized values.

Figure 4. Coefficient plot (full model) of zero-inflated negative binomial (ZINB) model

Figure 4 presents the coefficient plot for the zero-inflated negative binomial model. One may interpret the coefficients as follows: “For one unit change in the predictor, the difference in the logs of expected counts of the outcome variable is expected to change by the respective regression coefficient, given other predictors are held constant” (UCLA). The results in Figure 4 are to a large extent consistent with Figure 2. Crash counts are positively correlated with AADT, signalized intersections, and number of lanes in the major approach. Population density also has a positive effect on traffic crashes. Regarding other road-specific attributes, presence of median on approaches and lane widths of approaches show negative effects. Walk-miles traveled per capita increases crash counts, while speed limits at the approaches tend to decrease crash counts. Number of approaches arriving in the intersection has a positive effect. Other location features also show significant effects. Increase of distance to the nearest hospital reduces crash counts, but the presence of transit within 0.25 miles of the intersection centroid increases crash counts. As for census-level attributes, population density and average annual rainfall demonstrate positive effects, while median income and median age show negative effects.
Next, various tree-based ensemble ML models were used to predict crash occurrences at intersections across Texas, including random forest, extreme gradient boosting (XGBoost), light gradient boosting (Light GBM), and Bayesian additive regression trees (BART). The models had 42 features in total. This paper evaluated the performance of the models in the predictions. The procedures were as follows: (1) randomly split the data into 75% training and 25% test sets; (2) fit the model on the training data and generate predictions; and (3) evaluate model performance with various metrics, namely R-squared and root mean squared error (RMSE).

**Random Forest**

A random forest regression consists of decision trees generated by “splitting each node using the best among a subset of predictors randomly chosen at that node with a different bootstrap sample of the data” (Zhao et al., 2021). The random forest method computes the final prediction value based on the average results of each decision tree (Liaw and Wiener 2002; Li and Kockelman 2021). The number of trees was set to 500 in the random forest regression. This paper used the squared error to measure the quality of the split and considered all features when looking for the best split.

**XGBoost**

Chen and Guestrin (2016) devised the XGBoost method as a scalable ML system for gradient tree boosting. XGBoost constructs consecutive small trees with each tree correcting the net error from the previous trees (Zhao et al., 2021). XGBoost is trained in a forward “stage-wise” manner, aiming to minimize the sum of squared errors by tuning the parameters continuously (Li and Kockelman, 2021). “The first tree is split on the most predictive feature, and then the weights are updated to ensure that the subsequent tree splits on whichever feature allows it to correctly classify the data points that were misclassified in the initial tree. The next tree will then focus on correctly classifying errors from that tree, and so on. The final prediction is the weighted sum of all individual predictions” (Zhao et al., 2021). In the XGBoost training model, the maximum depth of the trees was set to 6, the number of rounds for boosting was 500, and eta (learning rate) was 0.1.

**Light GBM**

The Light GBM method is particularly useful for large datasets (Ke et al., 2017). It incorporates gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) (Li and Kockelman, 2021). The GOSS algorithm keeps all the instances with larger gradients while randomly dropping those instances with smaller gradients (Li and Kockelman, 2021). Light GBM speeds up the training process, thus reducing the computational time significantly. In the Light GBM model, the maximum number of leaves was set to 6, number of boosting iterations was 1000, and learning rate was 0.1.

**BART**

BART is a Bayesian non-parametric approach that fits a model using an influential prior distribution (Chipman et al., 2010). BART is a Bayesian “sum-of-tree” model in which “each tree is constrained by a regularization prior to be a weak learner” (Chipman et al., 2010). It performs iterative fitting and inference through conducting the back-fitting Monte Carlo Markov Chain (MCMC) that generates samples from a posterior. BART is robust to hyperparameter settings and addresses uncertainties with a Bayesian approach (Zhao et al., 2021). However, the method requires a lot of memory and time for computation. The number of trees was set to 100 for the model’s training.
COMPARISON OF MODEL PERFORMANCE

Balanced and Unbalanced Data

In the crash dataset, there were 522,933 (74%) zero-count intersections and 184,228 (26%) non-zero-count intersections. That made the data highly imbalanced. To address this issue, the dataset was resampled by implementing the `ovun.sample` function of the `ROSE` package in R, which is a "bootstrap-based technique that helps the task of binary classification in the presence of rare classes" (Lunardon et al., 2021). `Ovun.sample` generated synthetic balanced samples through a combination of randomly oversampling the minority class (intersections with non-zero crashes) and undersampling the majority class (intersections with zero crashes). In particular, it used bootstrapping to draw synthetic samples from the feature space neighborhood around the minority class to create new rows of new data for the minority class. It also randomly selected a set of majority class observations and removed those observations from the dataset (He and Garcia, 2009). After resampling, the numbers of zero-crash and non-zero crash intersections were approximately equal (zero crash: 353,813 and non-zero crash: 353,113), thus the balance of the dataset was adjusted. The modified sample was denoted as balanced data.

Signalized vs Unsignalized Intersections

Signalized intersections and AADTs exerted disproportionately high weights on the model predictions (This will be explained further in Section 5.4). As a result, other features were not well accounted for in the predictions. To deal with this problem, this paper subsetted the data into signalized intersections and unsignalized intersections. It also only included the intersections where the sum of AADTs of the incoming links exceeded 500 (i.e., excluding the low-volume sites). After subsetting the data, this paper found that there were 15,222 signalized intersections and 235,822 unsignalized intersections. Among the unsignalized intersections, 121,983 had zero crashes and 113,839 had non-zero crashes.

R-square and RMSE

Table 3 presents the summary of the model performances, in terms of R-square and root mean square error (RMSE). R-square and RMSE are commonly used metrics to evaluate model fit and performances for ML models (Li and Kockelman, 2021). Using the original (or imbalanced) data, we found that the R-squares of the ML models were not particularly high. The zero-inflated negative binomial model produced the worst predictions, as it yielded the lowest R-square and highest RMSE among all models. Concerning the five ML models, random forest regression resulted in the highest R-square (0.534) and XBART had the lowest R-square (0.508). The RMSE ranged from 10.64 to 19.71. Light GBM yielded the lowest RMSE, followed by random forest. The RMSEs indicated unsatisfactory predictions of the models. There were two possible reasons for this issue. First, the data contained a high proportion of zero-crash intersections. Second, there were a number of extreme values. For example, the maximum number of crashes was 996. Model predictions are likely to be affected by the extreme values. Resampling the data led to better predictions for some of the models. The R-squares increased across the models, with random forest reaching a R-square of above 0.8. RMSEs, on the other hand, only showed improvement for three models. RMSEs for the random forest, XGBoost, and ZINB models decreased by 2.07, 0.19, and 7.36, respectively, after resampling. Random forest’s RMSE saw the most significant improvement. However, the RMSEs for Light GBM and XBART increased by 1.92 and 7.82, respectively, which indicated poorer predictions for the two models, especially BART.
Table 1: Comparison of model performance: Imbalanced vs balanced data

<table>
<thead>
<tr>
<th></th>
<th>Imbalanced data</th>
<th>Balanced data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-square</td>
<td>RMSE</td>
</tr>
<tr>
<td>ZINB</td>
<td>-1.442</td>
<td>59.03</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.534</td>
<td>10.66</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.527</td>
<td>10.69</td>
</tr>
<tr>
<td>Light GBM</td>
<td>0.531</td>
<td>10.64</td>
</tr>
<tr>
<td>BART</td>
<td>0.508</td>
<td>19.71</td>
</tr>
</tbody>
</table>

Table 1 compares the performances of the ML models between signalized and unsignalized intersections. It shows that the R-squares are comparable across the two groups, but the RMSEs are much higher for signalized intersections. This is partly due to the higher variation of crash counts, in particular the higher number of extreme values, at signalized intersections. Unsignalized intersections had many more zero crash counts, thus yielding lower RMSEs that are comparable to Table 2. Considering model performances, one can see that the random forest model yielded the best model performance overall.

Table 2: Comparison of model performance: Signalized vs unsignalized intersections

<table>
<thead>
<tr>
<th></th>
<th>Signalized</th>
<th>Unsignalized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-square</td>
<td>RMSE</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.245</td>
<td>63.12</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.243</td>
<td>63.66</td>
</tr>
<tr>
<td>Light GBM</td>
<td>0.261</td>
<td>62.87</td>
</tr>
<tr>
<td>BART</td>
<td>0.203</td>
<td>83.69</td>
</tr>
</tbody>
</table>

Feature Importance

Given that random forest had the best model predictions in the analysis, this paper is interested in the relative importance each feature has on the predictions of the random forest model.

Figures 5a and 5b illustrate the feature importance of individual features using the imbalanced and balanced data, respectively. The graphs include the top 20 features in terms of importance. This paper scaled all measures of importance, such that the top feature had a maximum value of 100. The figures show that signalized intersections are the top feature, followed by AADTs, number of lanes of the approaches, and speed limit of the minor approach. Other important features included distance to the nearest hospital, distance to the nearest school, arterial minor approach, and walk-miles traveled. A number of census-tract level attributes were also important, including population density and median income. It is noteworthy that signalized intersections, and to some extent AADT at minor approach and AADT at major approach, had exceptionally high feature importance compared to other features. The three features exerted disproportionally high weights on the random forest model predictions.

As explained in previous section, the data was subsetted to focus on features other than signalized intersections and AADTs. Analyzing only the high-volume intersections where the sum of AADTs exceeded 500, Figures 6a and 6b illustrate the feature importance for signalized and unsignalized
intersections, respectively. They found that total walk-miles traveled, distance to the nearest school,
distance to the nearest hospital, population density, and employment density were the most important
features to the model predictions, although the five features were ranked differently between signalized and
unsignalized intersections.

Figure 5: Feature importance for the Random Forest models

(a) Feature importance for unbalanced data

(b) Feature importance for balanced data
Figure 6: Feature importance for the Random Forest models for signalized and unsignalized intersections with Sum (AADTs) ≥ 500

(a) Signalized intersections

(b) Unsignalized intersections
SENSITIVITY ANALYSIS OF CRASH PREDICTION

While regression and ML models excel at capturing relationships between features and outcome variables, the results may not be easy to interpret particularly for ML models. Specifically, one may find it difficult to quantify the substantive effects of each feature. Following Zhao et al. (2021), this paper employed a sensitivity analysis that captured the contribution each variable had on the model’s predictions. Let \( X \) be the set of features. The procedures of evaluating the sensitivity of variable \( X_i \) were as follows: (1) train the model on \( X \) and compute \( y \) as the prediction vector; (2) generate a new set \( X^* \) where a transformation is performed on variable \( X_i \); (3) generate prediction on \( X^* \) and define \( y^* \) as the prediction vector, and (4) compute the percentage change in the prediction mean, denoted as \( \frac{\mu^\prime - \mu}{\mu} \times 100\% \) (Zhao et al., 2021).

Following Li and Kockelman (2021), the transformation was as follows: (1) increase one standard deviation for continuous features; (2) binary change (0 to 1) for dichotomous features. Essentially, one standard deviation or binary change was implemented on each observation (Li and Kockelman, 2021). The new prediction was computed using the modified variables, and the difference between the mean of new predictions and original predictions represented the contribution of each feature (Zhao et al., 2021).

Figure 7 illustrates the sensitivities for the ZINB models and ML models for imbalanced and balanced data. Considering the more important features, number of lanes at the minor approach, speed limits at the major and minor approaches, and distance to the nearest hospital show different directions in Figures 10a and 10b. This was possibly due to the fact that different ML models interpreted the significance of the features differently (Zhao et al., 2021). Therefore, it is vital that one chooses the best performing model when one evaluates the metrics and examines feature importance with the optimum model. Since the ZINB model offers significance test of the predictor variables compared to ML models, this paper placed more weight on the results of the ZINB model when drawing inferences.

For the ZINB model, the road characteristics [local, collector, and arterial approaches] increased the outcome by a large percentage. In particular, arterial major approach had the most significant impact on total number of crashes. A binary change on arterial major approach could lead to a 214% increase in crash occurrences per intersection. The percentage changes for land use characteristics were smaller in the ML models. For example, a binary change on arterial major approach resulted in less than a 30% increase in crash counts for all ML models. For ZINB models, intersections in rural areas, small urban, and urbanized areas decreased crash counts by 137%, 59%, and 24%, respectively, compared to large urban areas. In the ML models, the percentage changes pointed to different directions for different urban-rural classifications. For XGBoost and BART models, rural areas decreased crash counts while for the random forest model, rural areas increased crash occurrences. Small urban and urbanized areas increased crash occurrences for most ML models. This contrasted with the ZINB results.

Regarding road design variables, the number of lanes and AADTs at the major and minor approaches had a significant impact on crash occurrence in the ZINB models. In the ZINB model, one standard deviation increase in the number of lanes at major approach led to a 59% increase in crash counts. In the random forest ML model, the percentage change decreased to 42%. In the ZINB model, one standard deviation increase in AADTs at major and minor approaches contributed to about a 144% and 52% increase in crash occurrences, respectively. In the ML models, AADTs at major and minor approaches also showed increases in crash counts. The percentage increases ranged from 35% to 108% on imbalanced data, and from 42% to 92% on balanced data. Signalized intersections also contributed to a large increase in the outcome for both ZINB and ML models. A binary change on signalized intersections contributed to 300% and 163% increases in crash counts in the random forest model on imbalanced and balanced data, respectively. As for census-tract level attributes, one standard deviation increase in population density, employment density, and...
precipitation increased crash counts by 33%, 20%, and 6%, respectively, while one standard deviation increase in median income and median age reduced crash occurrence by 50% and 43%, respectively. ML models showed the same directions in terms of percentage changes.

![Sensitivity for covariates for imbalanced data](image1)
(a) Sensitivity for covariates for imbalanced data

![Sensitivity for covariates for balanced data](image2)
(b) Sensitivity for covariates for balanced data

Figure 7: Sensitivity analysis for covariates in predicting total crash occurrence
(a) Sensitivity for covariates for high-volume signalized intersections

(b) Sensitivity for covariates for high-volume unsignalized intersections

Figure 8: Sensitivity analysis of covariates in predicting total crash occurrences for high-volume (sum (AADT) ≥ 500) signalized and unsignalized intersections
This paper compared the sensitivity analysis results of the ML models for signalized and unsignalized intersections in Figures 8a and 8b, respectively. Comparing the two figures, most features showed similar directions in percentage changes. Focusing on the most important features, we found that the percentage changes were mostly consistent across the ML models. Consider the top five features. When one standard deviation was increased to total walk-miles traveled, distance to the nearest school, population density, and employment density one at a time, this paper found positive percentage changes in crash occurrences for all models. The only exception was distance to the nearest hospital. Among signalized intersections, all models showed positive percentage changes except BART, whereas among unsignalized intersections, all models demonstrated negative percentage changes except random forest. It is noteworthy that random forest yielded relatively large percentage changes for the top five features.

CONCLUSION

This study presented a comparison between six different models- one econometric and five machine learning models- to explore the opportunity for machine learning models in predicting the motor-vehicle crash frequency and injury counts at the intersections across Texas. R-square and root mean square error (RMSE) metrics were used to evaluate the model fit and compare the model performances. Resampling of the data led to better prediction performances of all the models tested here and hence, the final comparison is made based on their performances on the balanced data. The zero-inflated negative binomial regression (ZINB) model was found to be the least accurate model in terms of both R-square (0.597) and RMSE (51.67). All the five machine learning models (ML) provided much higher prediction accuracy than the ZINBR model. The Random Forest model offered the highest prediction accuracy among the ML models with R-square value of 0.832 and RMSE value of 8.59 for the balanced data. XBARt model had the lowest prediction accuracy among the ML models with R-square 0.602 and RMSE 27.53 followed by Light GBM with R-square 0.647 and RMSE 12.56. Though resampling increased the prediction accuracy for all the models Random forest model saw the most significant improvement.

Neural network models were used to predict the total crash count and total crash count by severities of different intersections from the year 2010 to 2019 by aggregating the count by severity over 10 years. To improve the prediction accuracy, a deep neural network integrating a gated recurrent unit (GRU) is estimated for predicting total crash count by severities in the year 2019. The estimated model used temporal data from 2010 to 2018 as input to the GRU layer for capturing the temporal dependencies of traffic crash occurrences. This deep learning model showed higher prediction accuracy by decreasing RMSE by 26.6% and increasing R-square by 66% compared to the baseline neural network model for the total crash count by severities using the aggregated data.

While investigating the contributing factors of the crash events from the ML models the study found that signalized intersections and AADT both at minor and major approaches exerted disproportionately high weights on the model predictions. To deal with the problem this paper subsetted the data into signalized intersections and unsignalized intersections and considered only the intersections where the sum of AADTs of the incoming links exceeded 500. Both for signalized and unsignalized intersections, random forest (RF) model provided the highest accuracy in terms of both R-square (0.245 and 0.287) and RMSE (63.12 and 10.47) values. Analysis of the relative feature importance of the RF model for high-volume intersections (AADT > 500) showed that total walk-miles traveled, distance to the nearest school, distance to the nearest hospital, population density and employment density were the most important features to predict crash occurrence. Other important features included the number of lanes of the approaches, speed limit of the minor approach, arterial minor approach, and median income of the census tract.

Besides, the study carried out a sensitivity analysis to investigate traffic crash contributing factors by implementing one standard deviation increase (continuous features) and binary change (dichotomous features) for each observation. Sensitivity analysis showed that the effects of several variables have
different directions across different models making interpreting their contribution in predicting crash occurrences difficult. Since the ZINB model offers significance test of the predictor variables compared to ML models, this paper placed more weight on the results of the ZINB model when drawing inferences. The ZINB model showed that road types of the approaches (local, collector, and arterial approaches) increase crash frequency by a large percentage (214%) compared to the ML models. On the other hand, a binary change on the arterial major approach resulted in less than a 30% increase in crash counts for all ML models. For ZINB models, intersections in rural areas, small urban, and urbanized areas decreased crash counts by 137%, 59%, and 24%, respectively, compared to large urban areas. The percentage changes for different urban-rural classifications showed different directions in the ML models. For XGBoost and BART models, rural areas decreased crash counts while for the random forest model, rural areas increased crash occurrences. This made interpreting the influence of the predictors difficult and unreliable for the ML models. Among the road design variables, one standard deviation increase in the number of lanes and AADTs at the major and minor approaches significantly increases crash count in the ZINB model. In the ML models, an increase of AADTs also increased crash count, but an increase in the number of lanes led to a decrease in crash count contrasting the findings from ZINB model. Signalized intersections had been found to increase the crash count both in the ZINB and ML models. Among the census-tract level predictors, an increase in population density, employment density, and precipitation increased crash counts whereas the increase in median income and median age reduced crash occurrences. Both ZINB and ML models showed similar directions for the census-tract level variables.

Summing up, this study agrees with other similar studies (Iranaitalab and Khattak, 2017; Rahim and Hassan, 2021) upon the fact that machine learning models are better at predicting crash occurrences whereas statistical models are better at investigating the contributing factors of a crash event. The lack of test of significance and fluctuations of sensitivity of the predictor variables across models make the result ambiguous and unreliable. Different settings of the ML models may provide different results and change the drawn inferences. Traffic and transportation agencies can use ML models in predicting a crash event with higher accuracy, but care should be taken while investigating pre-crash conditions and influencing factors using ML models. Another important finding is that within the limited scope of this study recurrent neural network with GRU shows promises in dealing the spatial and temporal dimensions of traffic crashes. Future study can explore this potentiality in detail.

REFERENCES


Analysis & Prevention, 39(4), 657-670.


Rahman, M. S. (2018). Applying Machine Learning Techniques to Analyze the Pedestrian and Bicycle Crashes at the Macroscopic Level and Bicycle Crashes at the Macroscopic L.


